Reciprocal best hits are not a logically sufficient condition for orthology

Toby Johnson [1,*]

February 1, 2008

1. Département de Génétique Médicale
Université de Lausanne, CH

* email: toby.johnson@unil.ch

Summary

It is common to use reciprocal best hits, also known as a boomerang criterion, for determining orthology between sequences. The best hits may be found by BLAST, or by other more recently developed algorithms. Previous work seems to have assumed that reciprocal best hits is a sufficient but not necessary condition for orthology. In this article, I explain why reciprocal best hits cannot logically be a sufficient condition for orthology. If reciprocal best hits is neither sufficient nor necessary for orthology, it would seem worthwhile to examine further the logical foundations of some unsupervised algorithms that are used to identify orthologs.

Keywords: Reciprocal best hits, BLAST, orthologs

1 Introduction

Orthology between sequences means that they are vertically descended from a single common ancestral sequence [see e.g. Koonin 2005]. Many studies of molecular evolution rely on comparison of orthologous sequences from different species. With increasing amounts of sequence data available, increasing use is being made of unsupervised algorithms to identify orthologous sequences. A reciprocal best BLAST hit condition has been widely used to identify orthologs. Although this approach can be refined [e.g. Wall *et al.* 2003], the idea of a reciprocal best hit remains central to many methods for ortholog detection.

A reciprocal best hits method makes use of an algorithm, such as BLAST, that allows a *query sequence* to be queried against a *database* of sequences, and returns a ranked list of *hits*, which are sequences in the database that are similar to the query sequence. The top ranked hit is the *best hit*. Consider the task of identifying orthologs, when many sequences from each of two species are available. There are therefore two databases of sequences, with one database for each species. Then:

Definition 1 *Sequences s_1 and s_2, in databases G_1 and G_2 respectively, are said to be (pairwise) reciprocal best hits if:*

(i) s_2 is the best hit when s_1 is queried against G_2, and

(ii) s_1 is the best hit when s_2 is queried against G_1.

When s_1 and s_2 are reciprocal best hits, it is common to assume that they are orthologs. This is sometimes called the boomerang condition for (assumption of) orthology. The metaphor is that, if the condition is satisfied, one can start at s_1 (in G_1), go to s_1's best hit in G_2, which is s_2, and then go to s_2's best hit in G_1, and end up back where one started.

The notion of reciprocal best hits extends to three or more species. A natural definition is:

Definition 2 *Sequences s_1, s_2, …, s_m, in databases G_1, G_2, …, G_m respectively, are said to be (m-way) reciprocal best hits if:*

(i) s_j is the best hit when s_i is queried against G_j

for all $i, j \in \{1, 2, \ldots, m\}$.

If s_1, s_2, …, s_m are m-way reciprocal best hits, it would seem natural to assume that they are a set of m orthologs. When we have sequences from n species, some historically orthologous sequences may have been lost in some present day species. Therefore, we presumably do not wish to restrict ourselves to sets of $m = n$ orthologs, but will also be interested in finding sets of $m < n$ orthologs.

Wall *et al.* [2003] say that their reciprocal smallest distance method finds more sets of orthologs than reciprocal best BLAST hits, and that the sets of orthologs found by their method is a superset of those found by reciprocal best BLAST hits. This means that they consider reciprocal best BLAST hits to be a sufficient but not a necessary condition.

Poptsova and Gogarten [2007] state that the "reciprocal [best] BLAST hit method is very stringent and succeeds in the selection of conserved orthologs with a low false positive rate, but it often fails to assemble sets of orthologs in the presence of paralogs". This means that they also consider it to be a sufficient (with probability close to one) but not a necessary condition.

A key property of orthology, which stems directly from its biological definition, is that it is transitive. That is:

Property 1 (Transitive orthology) *If sequences s_1 and s_2 are orthologous, and s_2 and s_3 are orthologous, then s_1 and s_3 must also be orthologous.*

Note also that by definition:

Property 2 *Two different sequences from the same species cannot be orthologous.*

Most readers of this journal will be very familiar with everything I have said above. The purpose of this article is to point out the slightly surprising fact that definition 2 above, for reciprocal best hits, *logically cannot* be a sufficient condition for orthology. It contradicts properties 1 and 2. This result does not depend on any technical details of the algorithm used to find or rank the hits.

2 Theoretical Results

First, it is useful to note that definition 2 above, is equivalent to the following:

Definition 3 *Sequences s_1, s_2, ..., s_m, in databases G_1, G_2, ..., G_m respectively, are said to be (m-way) reciprocal best hits if:*

(i) s_i and s_j are pairwise reciprocal best hits, according to definition 1

for all pairs $i \neq j \in \{1, 2, \dots, m\}$.

The exact equivalence between definition 2 and definition 3 can be seen, by observing that definition 2 will be satisfied if definition 3 is satisfied, and *vice versa*. This equivalence is perhaps obvious, but it is worth emphasizing that definition 2 can be rewritten in terms of only pairwise relationships between sequences.

The problem with definition 2 (and therefore also with definition 3) can be explained concisely using a few terms from graph theory. In graph theory, a *graph* is an object that consists of *vertices* (or points), and *edges* (or lines) that connect some of the vertices. A *clique* is a part of the graph (i.e. a *subgraph*) for which there is an edge between every pair of vertices. We will mostly be interested in cliques that are not subgraphs of larger cliques, which are technically known as *maximal cliques*.

Consider a graph where every sequence, in every species, is represented by a vertex. Let there be an edge between two vertices if-and-only-if those two edges are pairwise reciprocal best hits. Then (using definition 3) sets of sequences that are (m-way) reciprocal best hits are the cliques of this graph. A well known property of cliques is that

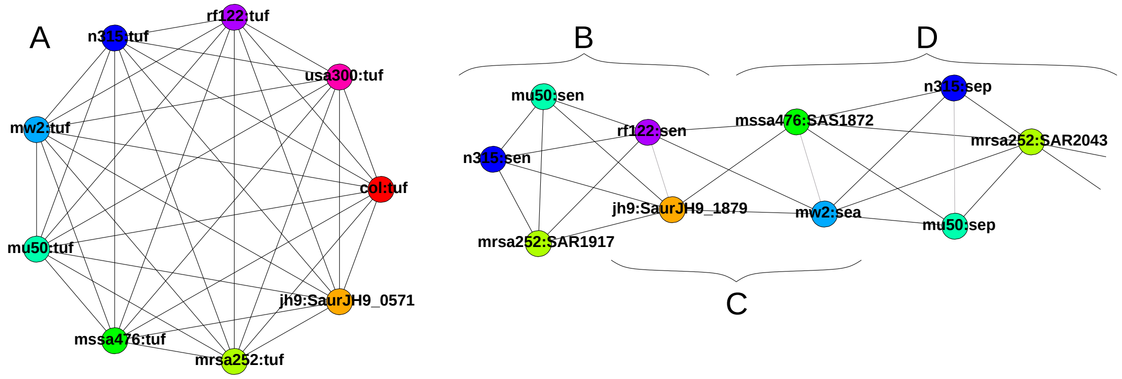

Figure 1: Subgraph of a reciprocal best BLASTP hit graph. Vertices represent coding sequences, are labelled by species:gene, and are colored according to species. Edges are drawn between vertices that are pairwise reciprocal best BLASTP hits. This subgraph contains one nonoverlapping maximal clique (A), and three distinct and overlapping maximal cliques (B–D).

it is possible for a single vertex to be a member of more than one maximal clique. The graph in figure 1 illustrates this property: it contains two five-member maximal cliques (labelled B and D), each of which has two sequences in common with a four-member maximal clique (labelled C). (None of the cliques B–D is a subgraph of any other, so they are three distinct maximal cliques.)

The example in figure 1 clearly illustrates why an m-way reciprocal best hit cannot be a sufficient condition for orthology. If all the members of clique B are orthologs, and also all the members of clique C are orthologs, then property 1 (transitive orthology) means that the set of sequences, belonging to either clique B or clique C, must all be orthologs. The same argument can be extended to clique D. We would be logically forced to conclude that, if reciprocal best hits is sufficient for orthology, then the set of sequences, belonging to any of cliques B–D, must all be orthologs. This contradicts the simple property 2 of orthologs, because it would mean that e.g. two sequences from the same species mu50, *sep* and *sen*, are orthologs. The fact that this *reductio ad absurdum* is possible, means that an m-way reciprocal best hit cannot be a logically sufficient condition for orthology.

The structure of the reciprocal best hits graph means that there cannot be an edge between two sequences from the same species, but there is no guarantee that longer range structures in the graph will be consistent with the assumption that cliques represent sets of orthologs. A graph representing true biological orthology, consistent with properties 1 and 2, consists only of completely separate cliques, each containing at most one sequence from each species, and with no edges connecting them to any sequence not in the clique. It might then seem that we can avoid the logical problem just described, by excluding cliques that have edges connecting them to any sequence not in the clique. Unfortunately, this cannot be a logically sufficient condition for orthology either. To see this, we can attempt to construct a meaningful definition:

Definition 4 (Perfect reciprocal best hits) *Sequences s_1, s_2, $\ldots$, s_m, in databases G_1, G_2, $\ldots$, G_m respectively, are said to be perfect (m-way) reciprocal best hits (with respect to other databases G_{m+1}, G_{m+2}, $\ldots$) if:*

(i) s_i and s_j are pairwise reciprocal best hits, according to definition 1

for all pairs $i \neq j \in \{1, 2, \ldots, m\}$, and

(ii) s_i and s_o are not pairwise reciprocal best hits, according to definition 1

for all $i \in \{1, 2, \ldots, m\}$ and for any s_o in any other database G_{m+1}, G_{m+2}, $\ldots$.

In graph theoretic terminology, sets of sequences that satisfy definition 4 are both (i) maximal cliques, and also (ii) *maximal connected subgraphs*.

Definition 4 cannot be a logically sufficient condition for orthology either, because it depends on the choice of "other databases", without which we cannot construct a meaningful definition of perfect reciprocal best hits. For example, consider again the subgraph in figure 1. If sequences from only three strains mw2, n315 and mu50 were analysed, the three sequences mw2:sea, n315:sep and mu50:sep would be perfect reciprocal best hits. However, if sequences from rf122 were also analysed, the same three sequences would not be perfect reciprocal best hits. Because the set of species actually analysed is determined by arbitrary choice and convenience, and because orthology is a biological property that is independent of which species are analysed, then perfect reciprocal best hits (definition 4) cannot be a logically sufficient condition for orthology.

3 Empirical Results

The mere fact that structures like the one in figure 1 are possible, means that technically speaking, reciprocal best hits cannot be a logically sufficient condition for orthology. However, it might be that this is rarely a practical issue. I therefore studied as an example, all coding sequences in nine strains of the bacterium *Staphylococcus aureus*. Whether these strains are considered to be distinct species is irrelevant to the argument being presented here, so I will use the terms species and strain interchangeably. I used all annotated coding sequences, and identified reciprocal best hits using BLASTP (i.e. querying protein sequence against protein sequence). These species are closely related, and so BLASTP is expected to be a reliable method for finding best hits.

The full graph, including all reciprocal best BLASTP hits, consists of 3411 maximal connected subgraphs. That is, there are 3411 smaller graphs, none of which are connected to each other. Of these, 3282 (96.2%) are also cliques, and therefore correspond to sets of sequences that are perfect reciprocal best hits. The size distribution of maximal connected subgraphs is given in table 1. Figure 2 shows the two largest subgraphs, each of size 22. Each can be seen to contain several cliques. Although only 2.8% of maximal connected subgraphs are not cliques, these include more of the larger subgraphs. In total, 5.9% of all sequences are in maximal connected subgraphs that are not cliques. Thus, we cannot determine orthology for over 5% of sequences using reciprocal best hits, and this proportion would have to (weakly) increase if more data from more species were included.

We can also quantify the extent of the problem by finding all cliques in the reciprocal best hits graph. There are 3793 cliques in total, and the size distribution of the cliques is given in table 2. Considering only cliques of size three or greater, 13.7% of cliques are not maximal connected subgraphs. This means that for 13.7% of sets of reciprocal best hit sequences, we should be cautious about inferring orthology, because some but not all sequences are reciprocal best hits with other sequences outside the set.

Thus, even in a closely related group of species/strains, there are reasonably common problems with using reciprocal best hits to determine orthology. It seems likely that the problems will be even more common with more distantly related species, or when attempting to identify orthologous regions of noncoding sequence in eukaryotes.

References

Koonin, E. V. (2005) Orthologs, paralogs, and evolutionary genomics. *Annual Review of Genetics* **39**:309–338. doi:10.1146/annurev.genet.39.073003.114725.

Poptsova, M. S. and J. P. Gogarten (2007) BranchClust: A phylogenetic algorithm for selecting gene families. *BMC Bioinformatics* **8**:120. doi:10.1186/1471-2105-8-120.

Wall, D. P., H. B. Fraser and A. E. Hirsh (2003) Detecting putative orthologs. *Bioinformatics* **19**:1710–1711.

Table 1: Size distribution of maximal connected subgraphs, for the reciprocal best BLASTP hit graph for nine *S. aureus* strains. The size distribution is broken down according to whether the subgraph is a clique or not; subgraphs of size greater than nine cannot be cliques.

Subgraph size	1–22	1	2	3	4	5	6	7	8	9
cliques (no.)	3282	581	154	118	77	58	54	72	196	1972
(%)	96.2	17	4.5	3.5	2.3	1.7	1.6	2.1	5.7	57.8
(% sequences)	94.1	2.5	1.3	1.5	1.3	1.2	1.4	2.2	6.7	76
non-cliques (no.)	129	0	0	2	6	7	7	6	11	16
(%)	3.8	0	0	0.1	0.2	0.2	0.2	0.2	0.3	0.5
(% sequences)	5.9	0	0	0	0.1	0.1	0.2	0.2	0.4	0.6

Subgraph size	10	11	12	13	14	15	16	17	18	19	20	21	22
cliques (no.)	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
non-cliques (no.)	6	16	15	7	6	5	10	2	2	1	1	1	2
(%)	0.2	0.5	0.4	0.2	0.2	0.1	0.3	0.1	0.1	0	0	0	0.1
(% sequences)	0.3	0.8	0.8	0.4	0.4	0.3	0.7	0.1	0.2	0.1	0.1	0.1	0.2

Table 2: Size distribution of cliques, for the reciprocal best BLASTP hit graph for nine *S. aureus* strains. The size distribution is broken down according to whether the clique is also a maximal connected subgraph (called "Perfect") or not (called "Imperfect"). Percentages are given relative to the total number of cliques of size 3 or greater.

Clique size	1	2	3	4	5	6	7	8	9	all
Perfect (no.)	581	154	118	77	58	54	72	196	1972	3282
(%)	NA	NA	4	2.6	2	1.8	2.4	6.6	66.8	86.3
Imperfect (no.)	0	107	94	80	82	75	40	33	0	511
(%)	NA	NA	3.2	2.7	2.8	2.5	1.4	1.1	0	13.7
Total	581	261	212	157	140	129	112	229	1972	3793

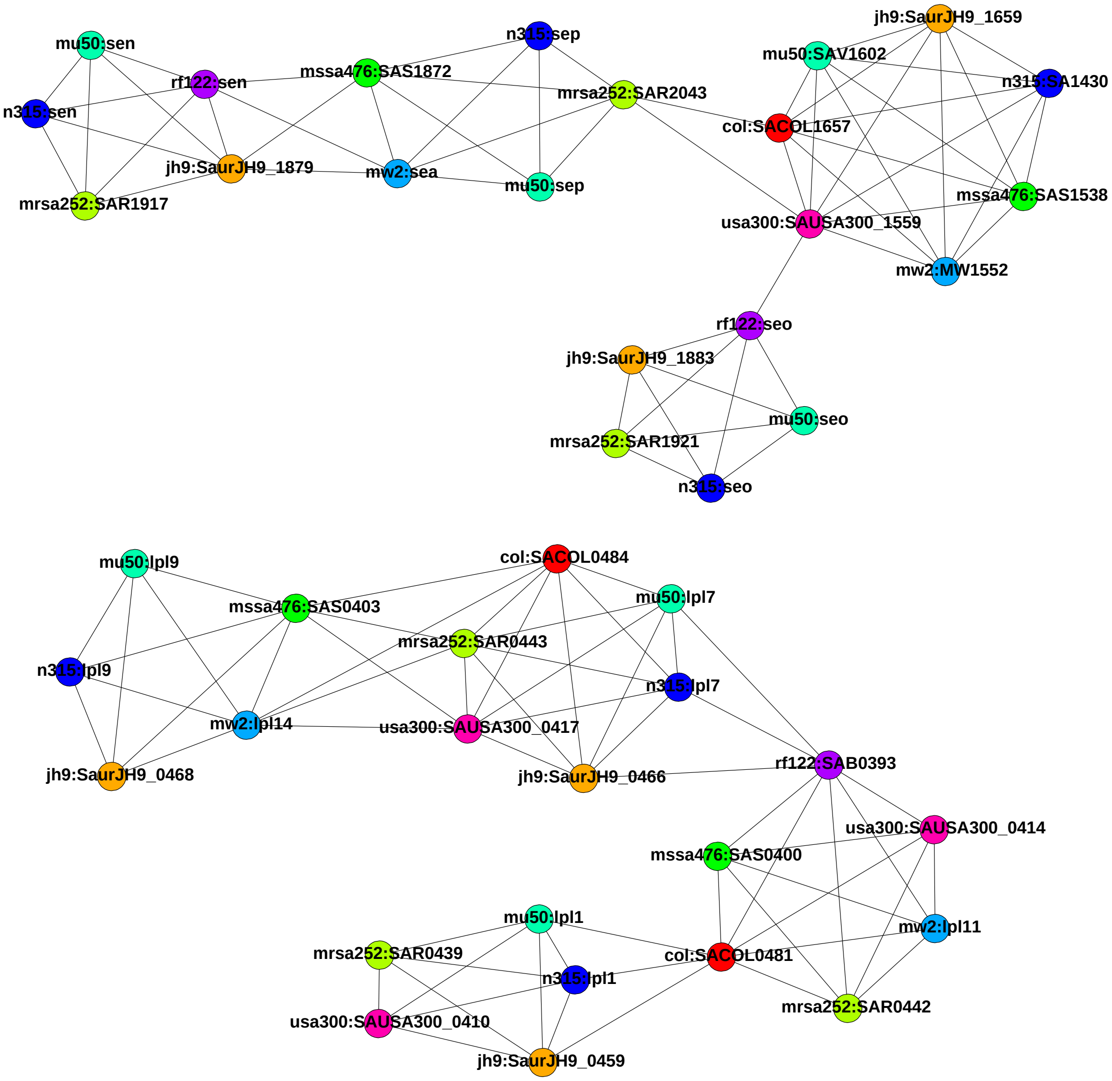

Figure 2: The two largest maximal connected subgraphs, each of size 22, of the reciprocal best BLASTP hit graph for nine *S. aureus* strains. Not surprisingly, these graphs contain an overrepresentation of coding sequences that have not been completely annotated.